NLP, Data Visualization, and the Mueller Report

Mike Mitri
CIS & BSAN
College of Business

Overview

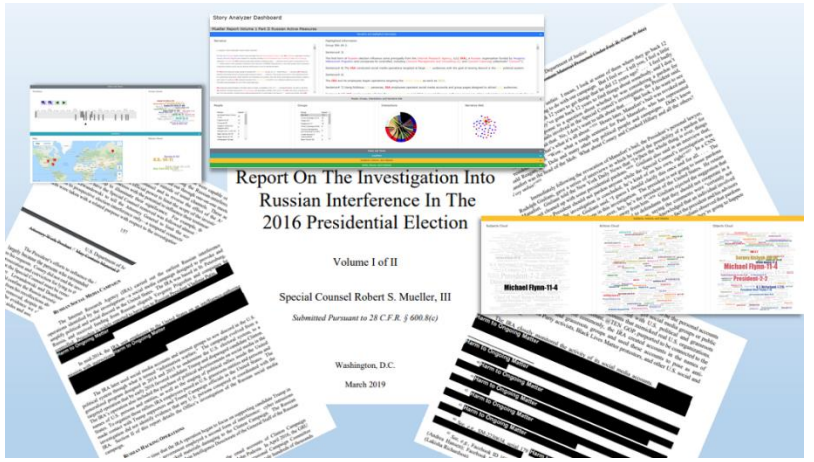
Story Analyzer http://storyanalyzer.org/

Natural Language Processing – Stanford's CoreNLP API

Information Extraction

Data Visualization – Data Driven Documents (d3) and Google

"A picture tells a thousand words"



Story Analyzer http://storyanalyzer.org/

An app that helps users visualize and understand a **story**

Story = a narrative

- People and groups are actors in the story. Actors can be subjects or objects.
- Subjects perform actions, objects receive the actions
- Times, places, and other contexts of actions
- Who did what to whom, where and when did it happen, and what else was going on at the time?

Story Analyzer uses NLP and data visualization APIs

Information extraction – take NLP results and capture narrative elements and relationships

Visualization – produce dashboards with visualizations depicting people, groups, interactions, times, and places

LinkedIn Articles

https://www.linkedin.com/posts/mike-mitri-a8912a1_activity-6580943797704810496-SAGg

https://www.linkedin.com/posts/mike-mitri-a8912a1_datavizualization-datastorytelling-activity-6586329975928999936-_es-

https://www.linkedin.com/posts/mike-mitri-a8912a1_nlp-congress-activity-6588102587873050624-hRV4

https://www.linkedin.com/posts/mike-mitri-a8912a1_nlp-datavizualization-datastorytelling-activity-6591323886409834496-W-nc

https://www.linkedin.com/posts/mike-mitri-a8912a1_nlp-storyanalyzer-informationextraction-activity-6595433941023420416-h953

Natural Language Processing (NLP)

A mix of artificial intelligence and computational linguistics

The study of "understanding" the natural human language

NLP involves grammar (syntax) and semantics (meaning)

What is “Understanding”?

- Human understands, what about computers?
- Natural language is vague, context driven
- True understanding requires extensive knowledge of a topic. Beyond the scope of NLP itself…requires ontologies.

http://en.wikipedia.org/wiki/Natural_language_processing

http://stanfordnlp.github.io/CoreNLP/ http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf

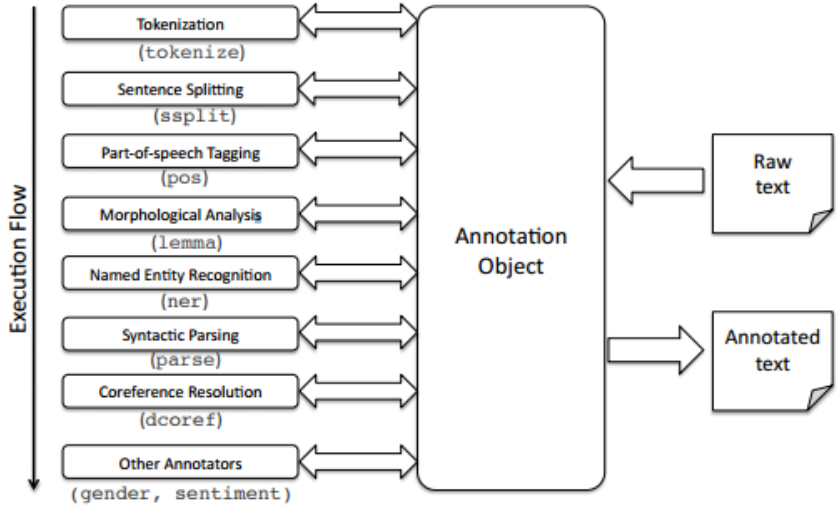
http://nlp.stanford.edu/software/

Stanford's CoreNLP

An open source Java-based API of classes and functions that can do several things:

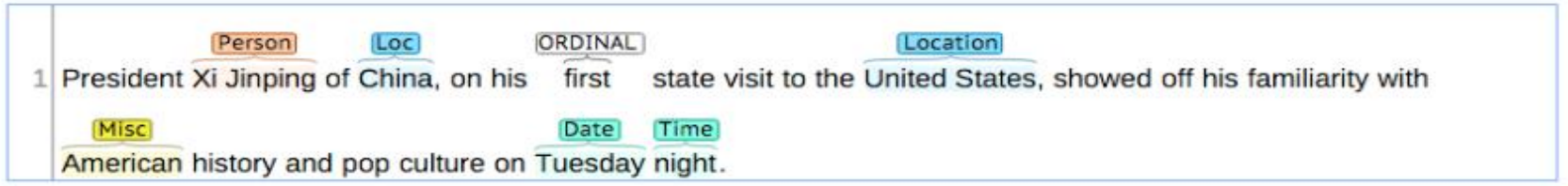
- Sentence splitting – breaking a text document into individual sentences
- Tokenizing a sentence (breaking it into individual "words")
- Identifying parts of speech (POS) within a sentence (nouns, verbs, adjectives, adverbs, etc.)
- Named entity recognition: Recognizing names of people, places, organizations
- Constituency parsing
- Dependency parsing
- Co-reference resolution – finding all expressions that refer to the same entity in a text. (e.g. finding connections between nouns and their associated pronouns)
- Temporal tagging – recognizing and normalizing temporal expressions

Stanford CoreNLP Features

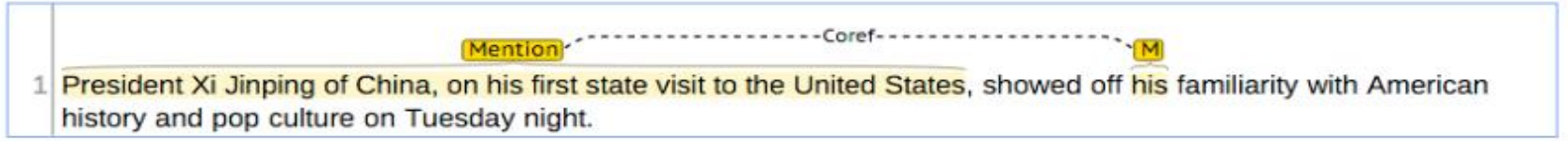


Some annotators based on machine learning, others rule-based.

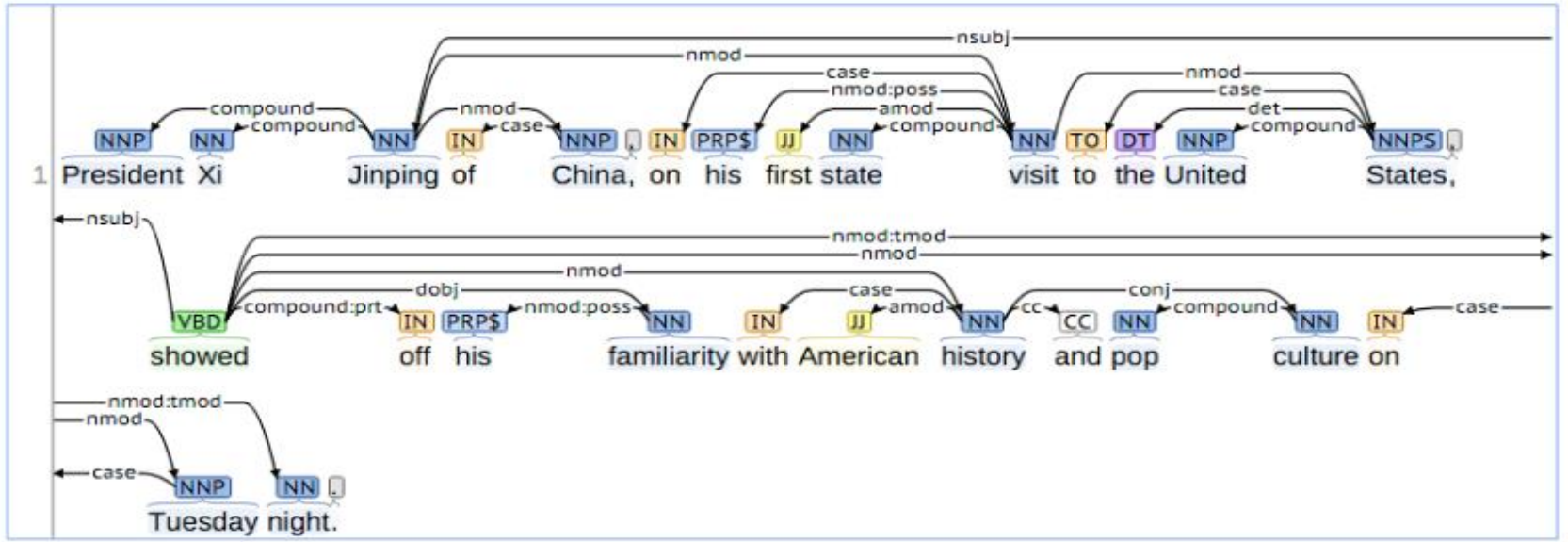
Named Entity Recognition:



Coreference:



Basic Dependencies:



Sentence Splitting

Sentences end in periods, question marks, or exclamation points.

But just because you have a period doesn't mean you are at the end of a sentence:

- Mr. Jones
- Samantha G. Jones
- Here is some text (i.e. something written).
- A, B, C, etc., etc., etc.
- Go to website cob.jmu.edu.
- What if the period is missing?

Tokenizing

Sentences are made up of words. Tokenizing splits up the sentence into its individual words.

At its simplest, tokenizing uses spaces as delimiters between words.

But, sometimes one word is actually a contraction of two:

- I'm, let's, isn't, won't

CoreNLP tokenizing also splits these into their individual constituents.

Punctuation characters are also tokens

POS Tagging

Once tokenized, the individual tokens can be recognized as **parts of speech**.

Stanford's CoreNLP uses the Penn Treebank Tag Set for recognizing parts of speech.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

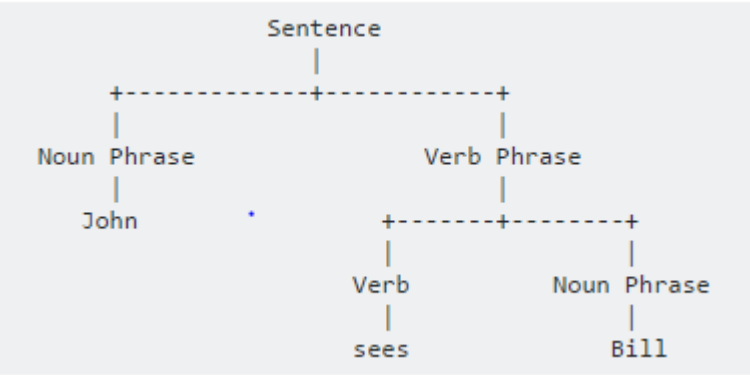
Named Entity Recognition

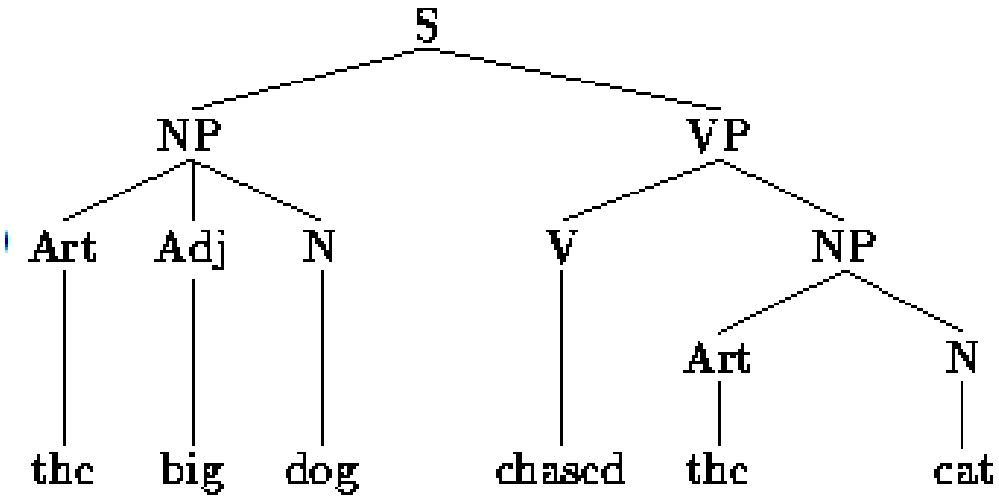
CoreNLP has models and classes for recognizing the names of:

- People
- Places
- Organizations
- Currency
- Time and date
- Extended version: Nationality, Religion, Ideology, Country, State/Province, City, and others.

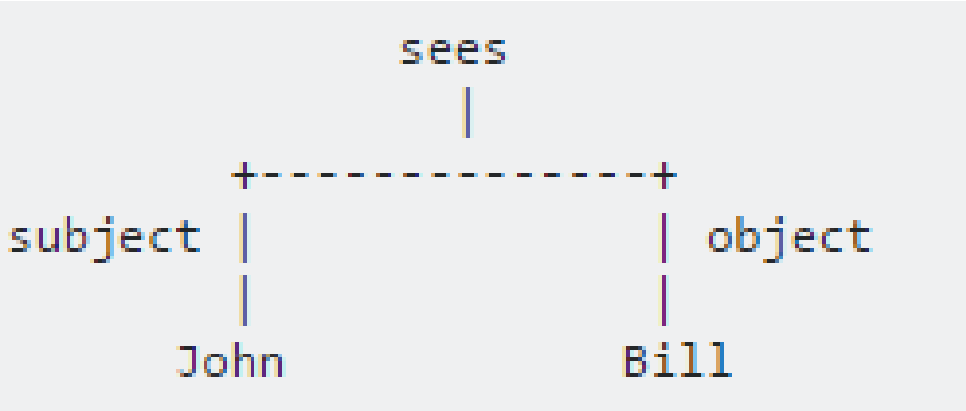
Constituency parsing

Hierarchy of phrases, sub-phrases, etc.





Dependency parsing



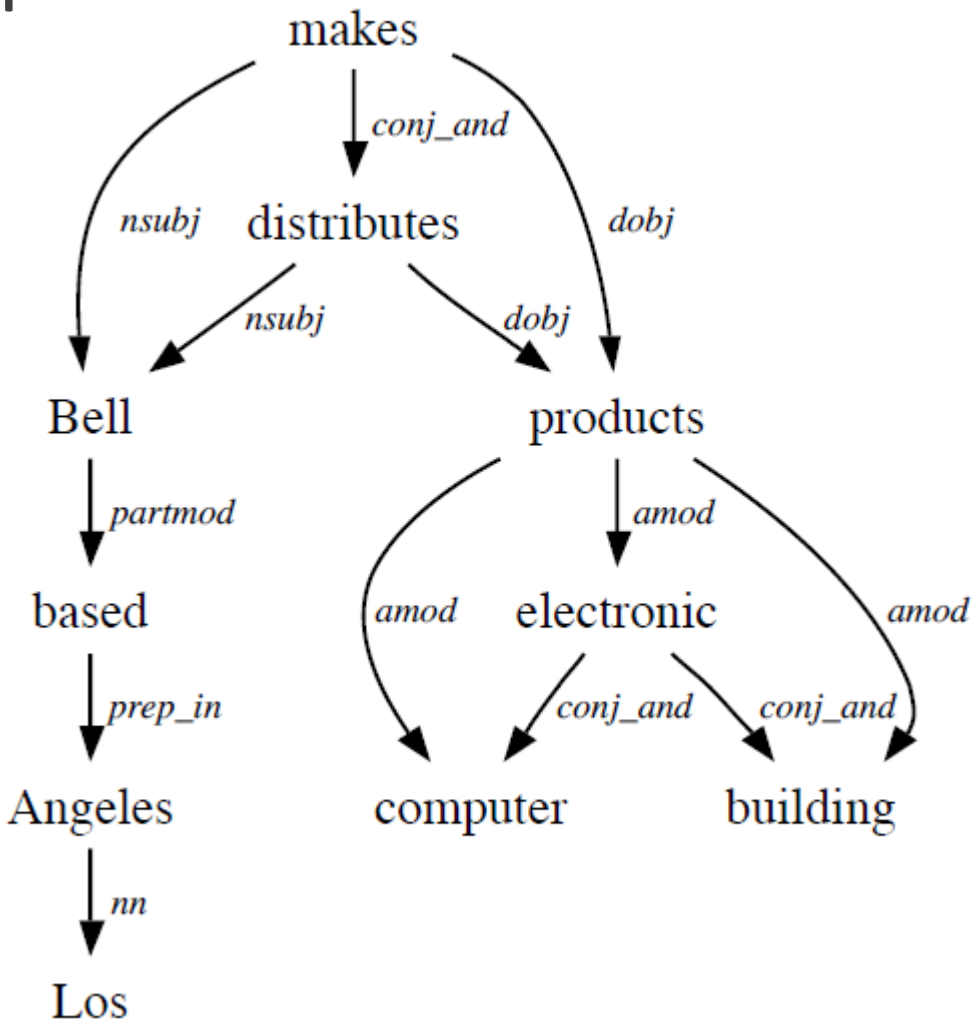
Dependency relationships (binary predicates) between words in a sentence.

In this picture, the **dependent** points to the **governor**

Example Dependency Graph

"Bell, based in Los Angeles, makes and distributes electronic, computer and building products."

From Stanford Typed Dependencies Manual (2008)
http://nlp.stanford.edu/software/dependencies_manual.pdf



Each connecting line is a dependency relationship. In this figure, the **governor** points to the **dependent**.

Dependencies from previous graph

nsubj – nominal subject

nsubjpass – nominal passive subject

dobj – direct object

amod – adjectival modifier

conj_and – conjoint and

prep – preposition (e.g. in, on, at, etc.)

nn – noun compound modifier

partmod – participial verb modifier

How to recognize subject-object relationships in text?

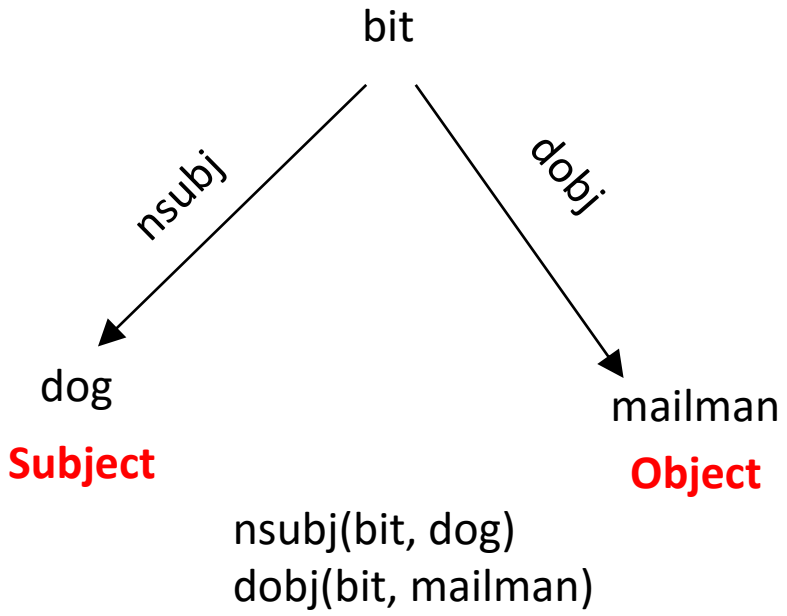
Subject-object relationships -- Who did what to whom?

- The dog bit the mailman
- The mailman was bit by the dog
- Trump beat Rubio in Florida, but he was defeated by Kasich in Ohio.
- The sun exerts gravity on the earth and on Mars.
- Hurricane Matthew bashes Florida with 100mph winds

Subject-object relationships

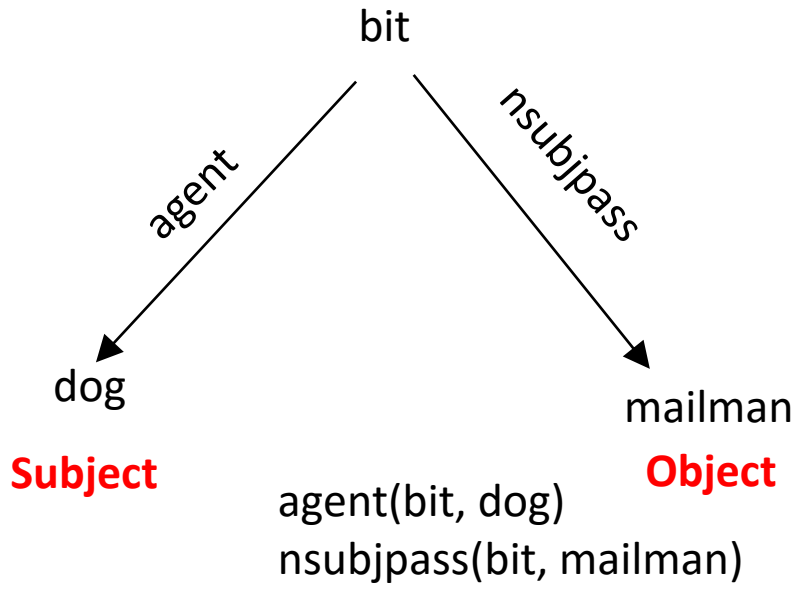
Active Voice

The dog bit the mailman.



Passive Voice

The mailman was bit by the dog.



Key subject-object dependency subgraphs in these sentences

Donald J. Trump beat[a] Marco Rubio in Florida, but then he[a] was defeated by John Kasich[a] in Ohio. Later, though, he[b] beat[b] Kasich[b] and all others at the convention.

- nsubj(beat[a],Trump), dobj(beat[a],Rubio)
- nsubjpass(defeated,he[a]), agent(defeated,Kasich[a])
- nsubj(beat[b],he[b]), dobj(beat[b],Kasich[b])

Note: there are six entities (subjects or objects) cited: Trump, Rubio, he[a], Kasich[a], he[b], and Kasich[b]

Coreference Resolution

Identification of **coreference chains**

A coreference chain has a list of **mentions**

Each mention refers to a word (or cluster of words) in the text

- Mention type – list, nominal, pronominal, proper
- Gender – male, female, or neutral
- Animacy – animate or inanimate

Coreference Resolution

Donald J. Trump beat[a] Marco Rubio in Florida, but then he[a] was defeated by John Kasich[a] in Ohio. Later, though, he[b] beat[b] Kasich[b] and all others at the convention.

What are the **coreference chains**?

1. Trump and he[a] and he[b]
2. Kasich[a] and Kasich[b]

Resulting dependencies after **coreference resolution**. Results in three entities (Trump, Rubio, and Kasich[a]):

- nsubj(beat[a],Trump), dobj(beat[a],Rubio)
- nsubjpass(defeated, Trump), agent(defeated,Kasich[a])
- nsubj(beat[b], Trump), dobj(beat[b], Kasich[a])

Limitations of NLP

Accuracy measures of CoreNLP annotators.

CoreNLP Annotator	F1 Score	Test Information Source
POS Tagging	97	https://nlp.stanford.edu/software/pos-tagger-faq.html
Dependency Parsing	81	https://nlp.stanford.edu/software/stanford-dependencies.shtml
Named Entity Recognition	81	https://nlp.stanford.edu/software/crf-faq.shtml
Coreference Resolution (NN)	60	https://stanfordnlp.github.io/CoreNLP/coref.html

What is an F1 Score?

In machine learning, an F1 score is a metric for measuring accuracy in machine learning (data mining)

Machine learning model will make a prediction. For example prediction of whether a customer will buy a product (simple yes/no). The accuracy of the model is how reliable the prediction is.

F1 is a score between 0 and 100

Two components of F1:

- **Precision** = What percentage of times did the model predicted yes and the customer actually bought a product.
- **Recall** = What percentage of times that the customer actually bought a product did the model predict this?

There is often a tradeoff between precision and recall. Think about law enforcement.

Winograd Schema Challenge (WSC)

Hector Levesque – Winograd Schema Challenge

Common sense reasoning, alternative to the Turing Test

Example: pronoun resolution

“The trophy couldn’t fit in the briefcase because it was too big.”

vs.

“The trophy couldn’t fit in the briefcase because it was too small.”

Coreference resolution’s poor performance shows that computers have a long way to go to pass the WSC.

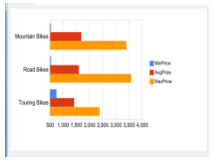
Data Visualization for NLP

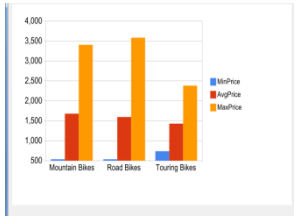
To recap, NLP gives us this:

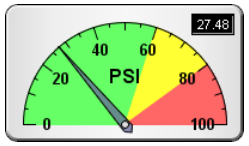
- ❑Individual sentences with individual tokens (including words)
- ❑Parts of speech for each word
- ❑Binary dependencies between pairs of words
- ❑Name recognition of people, groups, places, times, etc.
- ❑Coreference chains involving mentions (with a head mention)
- ❑Story Analyzer creates ASO instances

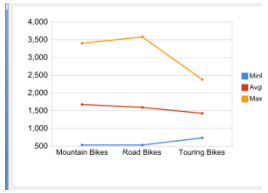
What visualizations are good for capturing and displaying this data structure and content?

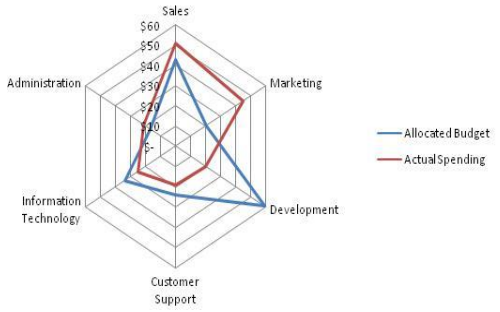
Common Visualizations found in many Dashboards

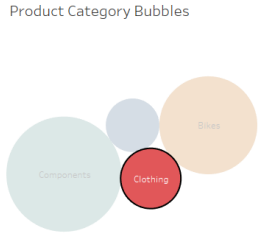


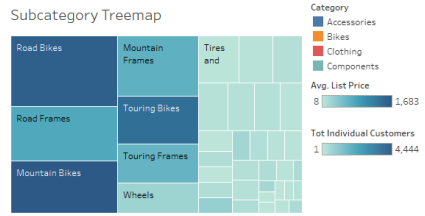


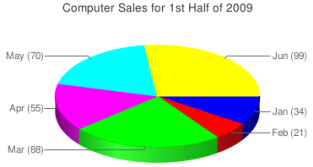




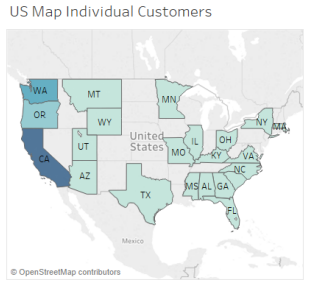




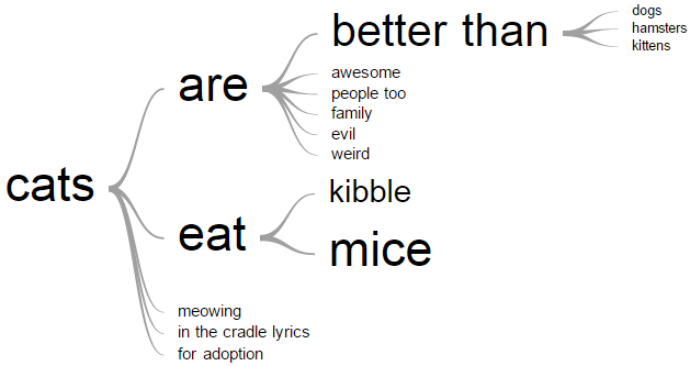




These depict numerical and categorical data – Structured Data



Visualizations for NLP Results and Concepts



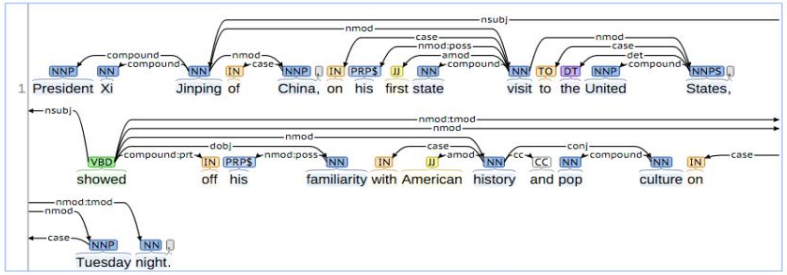
Named Entity Recognition:

1 President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

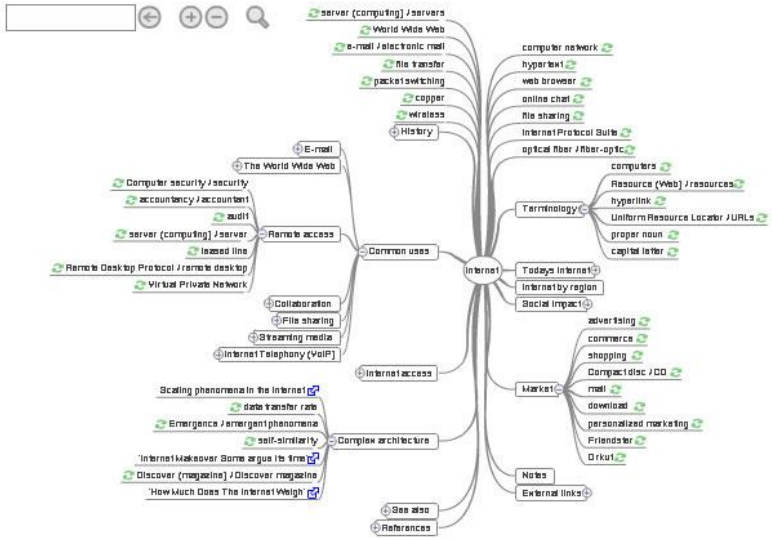
Coreference:

1 President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

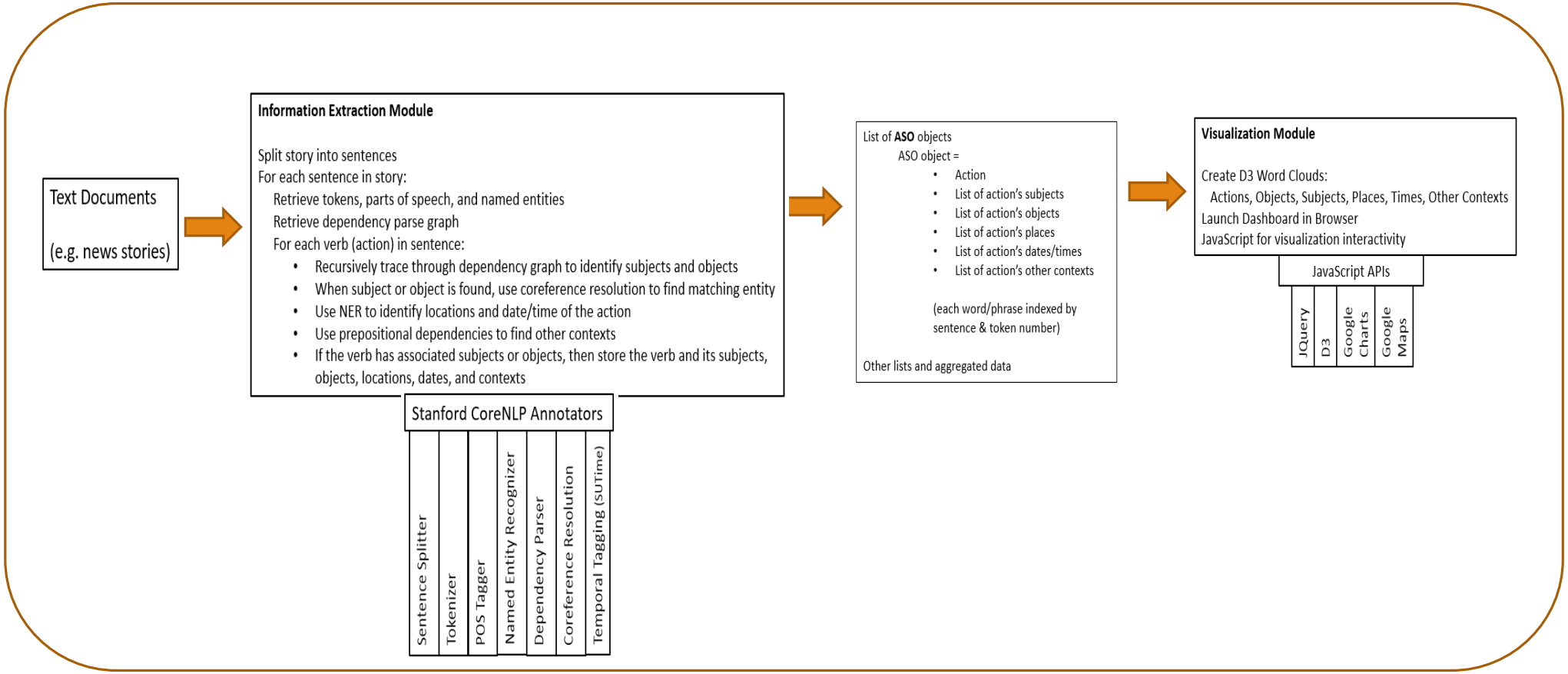
Basic Dependencies:



NLP takes “unstructured data” (raw text) and imposes a structure



Story Analyzer Architecture



Information Extraction

Extracting structured information from unstructured text with help from NLP

Often called “slot-filling”

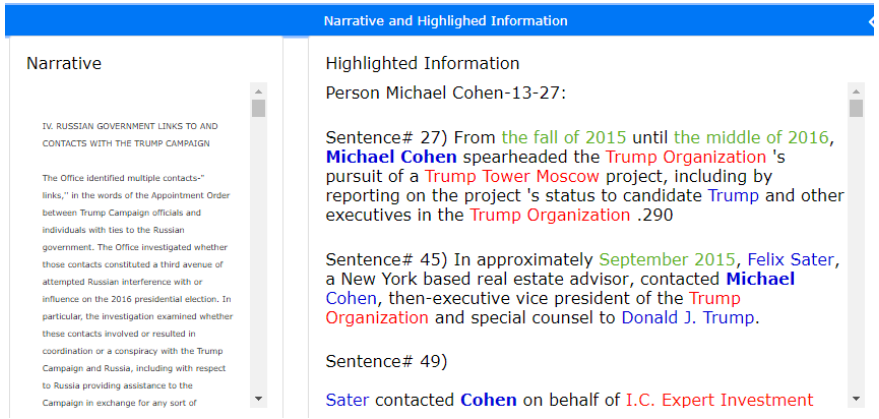
Story Analyzer structure builds on CoreNLP structure, relies heavily on named entity recognition (to recognize people, groups, etc.), dependency parsing (to recognize interactions), and coreference resolution (to identify the head mention)

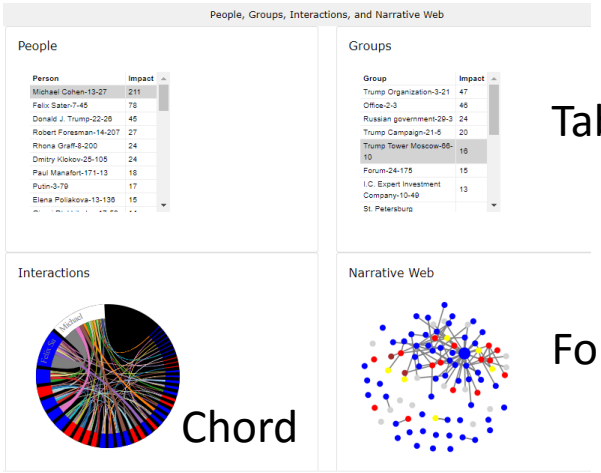
Action/Subject/Object (ASO) instances

- Each action (verb) will have
  - List of subjects
  - List of objects
  - List of places
  - List of times
  - List of other contexts

Story Analyzer Dashboards have Six Sections

The six sections





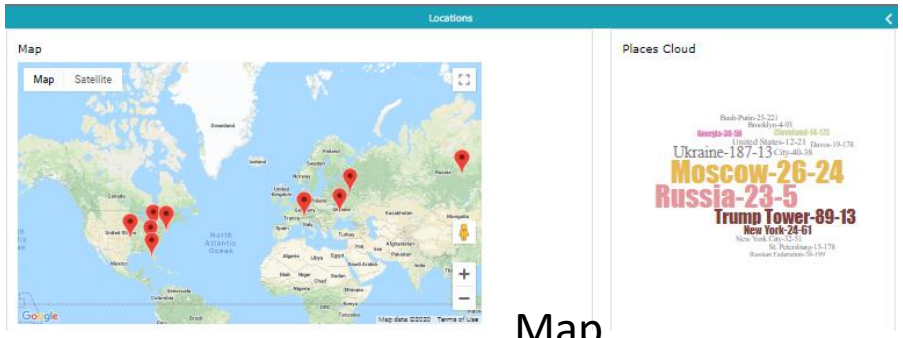
Tables

Chord

Force graph

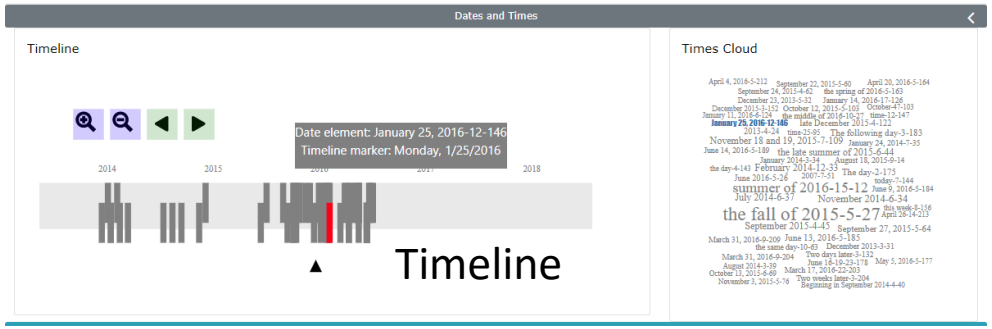
Visualization APIs:

- d3 (data driven documents)
- Google

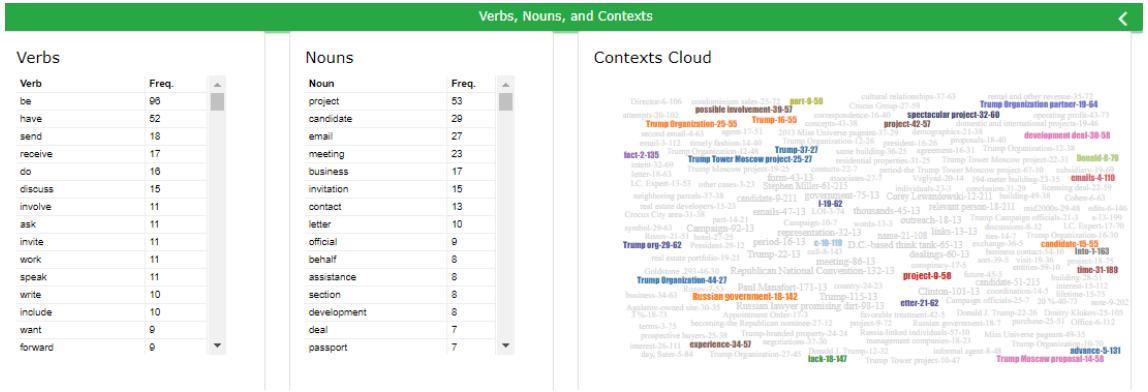


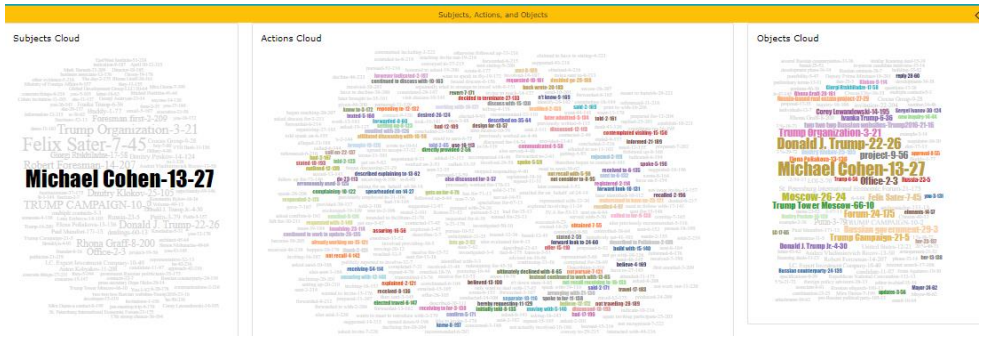
Map

Interrelated word clouds



Timeline

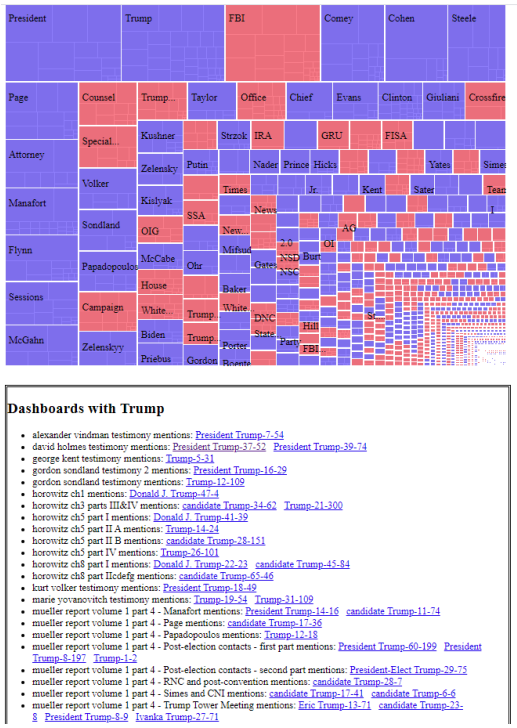


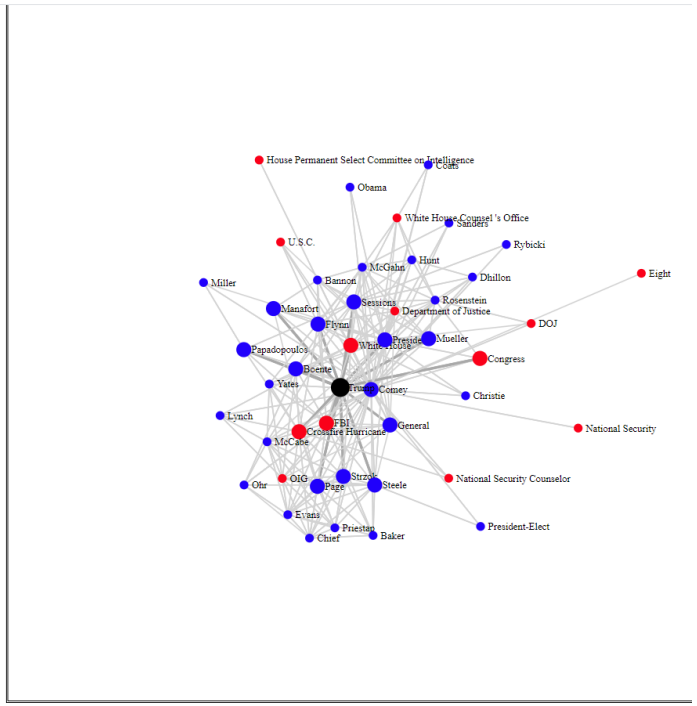


Getting the Big Picture

A separate dashboard:
http://storyanalyzer.org/overviewdashboard.html

Presents a TreeMap visualization. Can see people and groups, and their interactions, across multiple documents. Clicking a name produces a force graph depicting interactions with other people or groups. Hovering over a node shows documents and mentions. Clicking a mention takes you to the specific dashboard.





Applied to several Official Governmental Documents related to Impeachment

Mueller Report

House impeachment testimony opening statements

Republican Talking Points

Horowitz report

House Impeachment Report

White House Response

Senate Acquittal Report?

Currently 50+ dashboards. Final estimate for this project: 100+

Dashboard Generation Procedure

Can be done by student assistants:

Clean the text

Run against Story Analyzer for NLP processing

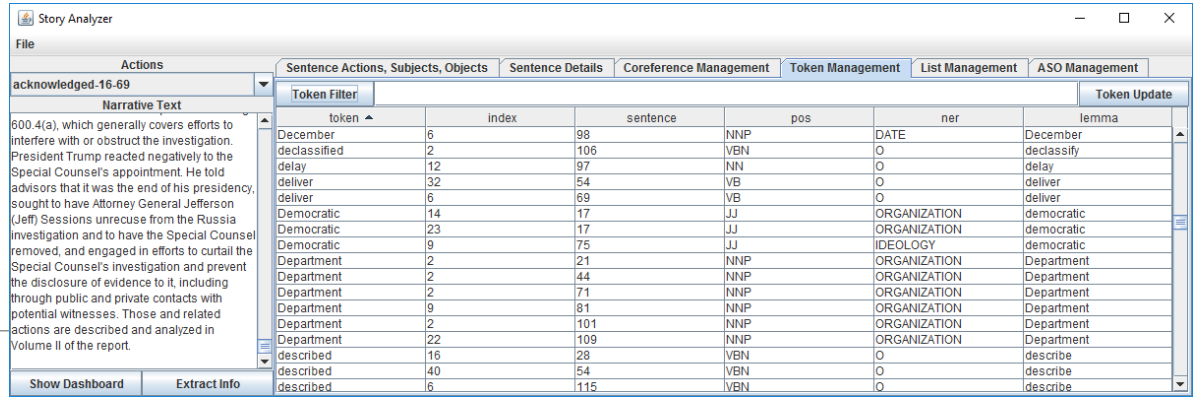
Use Story Analyzer Editor to edit NLP results and correct NLP errors

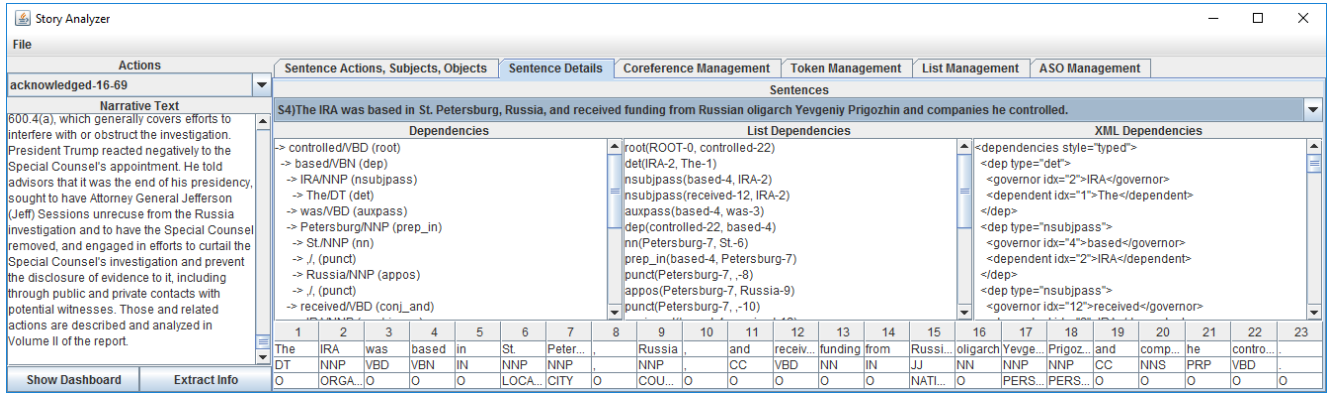
Run Information Extraction (producing A/S/O and other StoryAnalyzer data)

Run Dashboard data generation (json data)

Copy data to dashboard templates and json repository

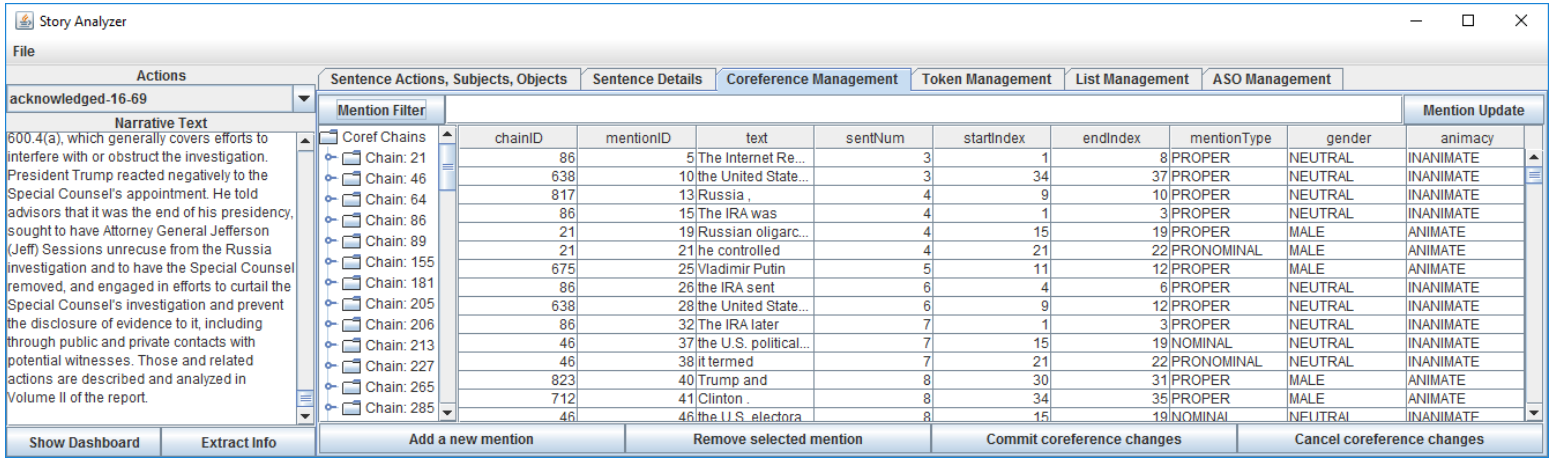
Editing NLP Results





Correcting and refining NLP results.

Information extraction and dashboard data generation.



Conclusions and Future Research

Applications to other areas:

- Health care – health claim fraud detection
- News – enhanced newsreader experience
- Information Systems discipline – User stories
- Law – Analyzing legal documents
- Business --

Improvements and enhancements

- Refine information extraction algorithm
- Incorporate ontologies and “common sense”
- Experiment with more visualization techniques